International Journal of Mechanical and Production Engineering Research and Development (IJMPERD)
ISSN (P): 2249-6890; ISSN (E): 2249-8001
Vol. 8, Issue 3, Jun 2018, 275-284

A BIG DATA MODEL FOR AN INFERENCE ENGINE OF A PRODUCTION COMPANY

MANISHANKAR. S[1] & S. SATHAYANARAYANAN[2]

[1]*Research Scholar, Department of Computer Science, School of Computer Science & Engineering, Bharathiar University, Coimbatore, Tamil Nadu, India*

[2]*Research Supervisor, Department of Computer Science, School of Computer Science & Engineering, Bharathiar University, Coimbatore, Tamil Nadu, India*

ABSTRACT

Recent trends in data processing and analytic systems prove that the amount of data generated in production companies are massive and handling them is a herculean task. Many data centers of such companies are currently facing efficiency dearth due to the increasing demand in the data analytics and processing. Cloud- based data centers are now adapted to the processing technology developed by platforms like Hadoop. Hadoop has provided a solution for high scale data processing with Map-Reduce Algorithm as the root. Data analytics is considered a vital aspect of Big Data handling. There are multiple models of Hadoop analytics available. In this paper, a multi- model analytics approach is proposed for the performance improvement of Hadoop. The model improves the dimension by performing a procedure of categorization of data and identification of missing values. With the help of this technical paper put forwards hybrid model of Hadoop analytics that is a combination of Hadoop with K-Mean Clustering Analytics and Hadoop with EM algorithm Analytics according to the data results of preprocessing. The model gives some faster analytics for Hadoop processed data which can result in improvement of an analytical system in a production company.

KEYWORDS: R, Big Data, K-Means, Hadoop, EM-Algorithm, Handling Missing Data & Performance Evaluation

Received: Mar 15, 2018**; Accepted:** Apr 05, 2018**; Published:** Apr 30, 2018**; Paper Id.:** IJMPERDJUN201831

Original Article

INTRODUCTION

In recent years there is a huge increase in data that is generated from various sources which are collected and stored in a data warehouses or data centers [1]. It is definitely a challenging task for any enterprise to store these huge data, analyze and process it. Theses humongous data-based knowledge discovery referred to Big Data mining has the ability to extract information which is useful [2]. These datasets will extend the boundaries of traditional data processing, analyzing and storing the data and is given as Big Data processing and analytics[3]. Big Data can be classified as three' V's they are Volume describing the size of the data, Velocity is where the data will be developed and Variability is the complexity of data and its structure how the data is interpreted [4].

To solve the difficulty in big data processing Hadoop Map-Reduce platform is available, which is an open source environment[5]. Hadoop provides distributed processing platform which does data processing in a large-scale manner. The available data is processed using Hadoop using a specific Map-Reduce function which will provide high performance. The Map-Reduce function consists of two functions a map function and a reduce function. The map function is written in such a way that multiple map functions can be performed at once, so the program parts will be divided into tasks[6]. The set of data will be will be converted into another set of data where

the elements will be broken into tuples. The reduce function will take map function output, and process it, which combines the values which create the result output file this kind of outputs are used in various analytical application and results [7].

RELATED WORK IN BIG DATA ANALYTICS AND PROCESSING

Day by day increase data arises a need for various data analytic professionals to study on various aspects of the Big Data Mainly three dimensions, categorization, processing and analytics. In Scalable Random Forest (SMRF) algorithm there is an improvement of traditional random forest algorithm based on Map-Reduce. It helps to work in parallel processing of huge data in distributed computer cluster and the classification for huge datasets. This SMRF algorithm will work well and it will convert the distributed computing environment to choose the tree scale. So here we can come across that SMRF algorithm is better than the traditional Random Forest algorithm as the performance will be improved [8]. Hadoop is introduced for big data which can access vast data. Where Hadoop uses Map-Reduce to process data. Map-Reduce is for large-scale parallel processing, analytics for processing massive data. When Hadoop runs Map-Reduce program, it will send the job to master node and job tracker has many slave nodes which will be assigned to a new work when it is idle [9].

The huge need to be pre-processed using any processing techniques to remove the duplicate data. The words are handled which gives an improvement in accuracy [10]. Big data cleansing using data auto-discovery of quality rules and some conditional function dependency. Here they are using big data pre-processing quality framework which will help them to solve the numerous data quality concerns and that occurs when attempting to apply quality concepts to huge data [11] the performance of data processing using Hadoop Map-Reduce in the first party they are explaining about Hadoop Map-Reduce and HDFS in details and they will contrast both with parallel database. The static physical execution plan of Hadoop Map-Reduce and we can check the effects of job performance [12]. Big data in the context of cloud computing. Here we discuss about key issues, including cloud storage computing architecture, popular parallel processing framework for the optimization of map-reduce [13].

Big R is the new platform which enables retrieving, handling, evaluating, and visualizing the data existing on a Hadoop cluster from R user interface. Big R is an R package which overloads a number of functions in R and operators to be able to access big data on a Hadoop cluster [14]. The organizations can predict the output by analyzing big data. There are many data which are unstructured which are suitable for data mining and subsequent analysis and provides flexible, low cost, many business problems will be solved by reducing the risk, and satisfy customers' needs and will get profit [15]. Here we can come across the developed algorithm to compute logistic regression using gradient ascent. Gradient ascent algorithm will help in Map-Reduce programming and also in large-scale data analysis. Here we can come across the implementation of R Hadoop and RHIPE and spark R and they have done performance study of different R packages. It checks for their runtime, scalability, and usability. This implementation will improve the scale of the data that can be analyzed with R. All this is parallel execution which will improve the response time and processing time of the system [16].

PROPOSED MODEL OF IMPROVED DATA ANALYTICS

A glance through the literature work clearly denotes that the work carried out to make some efficient analytics for the Big Data is still in a budding phase. Many of the techniques used have some inherent data related issues. The research carried out in this paper proposes a multistage analytical platform which overcomes the data related issues. The initial

phase starts with the data collection from distributed storage server and then data processing has to begin. The problem arises here because normally Hadoop map-reduce will consider all data in the same way and process it. So later analytics becomes difficult to perform. So, an engine is proposed which performs initially the only categorization of data according to type and features and then chooses the processing and analytics platform. Here we consider some Big Data collected from an intuition library. So here the data contains some portion of invoice details which is large and outnumbered. As well as there are some data which is sparse, and which contains missing parts are there in the data set. So the learning machine will push the portion of the data which will thick and outnumbered is forwarded in to Hadoop with K-Means analytic platform or if it is sparse and missing data it will be pushed to Hadoop with EM algorithm. Thus solving the problem of data dissimilarity and hiccup in later stages of Hadoop processing and Analytics.

ARCHITECTURE DIAGRAM

The proposed system is a combination of supervised and unsupervised learning. The architecture diagram depicted portrays the various learning modules which carries out the Big Data processing and analytics for the given data in much faster and efficient manner. The first component is the data collection part which receives from various distributed devices like web applications, mobile applications, cloud-based applications etc. which produce huge data. Data learning machine is the next part which performs categorization of data in to thick or sparse based on a supervised machine learning algorithm built up with R. This will in turn, suggest the processing platform which has two parallel platforms first one with Hadoop with K-Mean which take the thick portion of the big data and second one is Hadoop with EM (Expectation Maximization) which handles the sparse data. Thus the information system will get the processed and analyzed data needed to work in efficient and faster manner. Thus making a better Data processing and analytical system

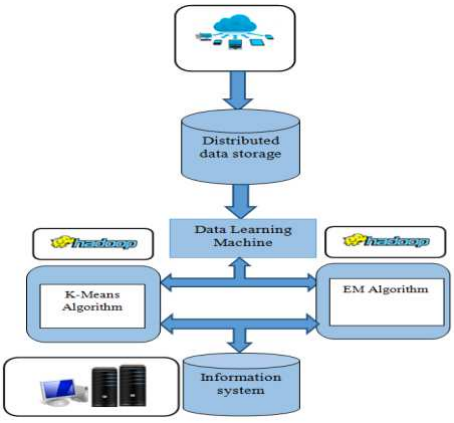

Figure 1: Architecture Diagram

CATEGORISATION OF DATA AND IDENTIFICATION OF MISSING VALUES

Initial process done by Data Learning machine in the proposed system is to categorize the data so that it can be processed efficiently using Hadoop. A major function in the pre-processing analytics involves finding sparse and thick data and as well as dealing with missing values that occur frequently as the data size increase. So here we use an algorithm with p Miss function to identify sparsely and identifying the number of missing fields in data. Mice package in R which helps you assigning missing values. There are two types in missing values. MCAR- Missing Completely at Random. The tendency for missing data is completely random. MNAR- Missing Not At Random. The tendency for missing data is not linked to missing data, but to some of the observed data.

Algorithm for Missing Data

Pmm-Predictive mean matching

- Take the input and apply missing value function.
- pMiss<-function(x){sum(is. a(x))/length(x)*100} //calculate the missing values
- apply(library,2, Miss)

  //apply function to dataset.

1. Import mice package.
2. library(mice)

   //Extract mice package from library
3. md. attern(library)

   //check the pattern of missing data

RESULTS

Sl. No	Emp ID	Employee	Machine	Production Unit
0.0000000	0.0000000	0.0000000	0.7290899	0.0000000
Time	Shift	Hours	Wage	
47.6202179	0.4341771	0.4505612	0.0000000	

Table 1: Summary of EM cluster

	Sl. No	Employee	Machine	Production Unit	Time
5	1	1	1	1	1
12147	1	1	1	1	0
55	1	1	1	0	0
	0	0	0	55	12207
Shift	Hour	Wage	Leave	skill	
0	0	0	0	4	
0	0	0	0	5	
0	0	0	0	6	
12207	12207	12207	12207	61085	

1. library(VIM)

   //Extract VIM package from library
2. aggr_plot<-aggr(library, ol=c('navyblue','red'), umbers=TRUE, sortVars=TRUE, labels=names(data), cex. xis=.7, gap=3, ylab=c("missing data", “Library data"))

   //Graph plotting for the missing

   values.

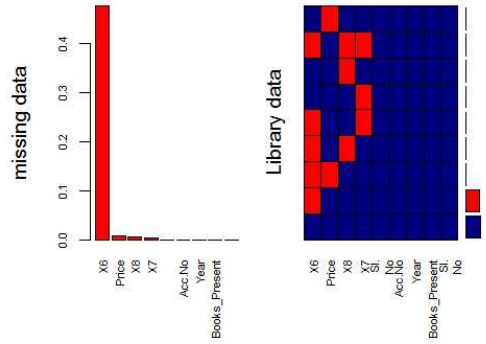

Figure 2: The Result of the Missing Values

marginplot(library[c(1,2)])

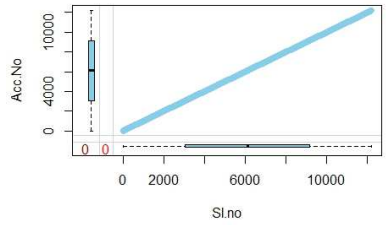

Figure 3: Plotting Two Variables at a Time where Blue Plot is the Remaining Data Point

1. tempData <- mice(library, =1, axit=10, eth='pmm', eed=500)

Missing cells per column

Table 2: Summary of EM cluster

Sl. No	Employee. No	Employee ID	Machine	
0	0	0	89	
time	Hour	Wage	Shift	Skill
0	5813	53	55	0

2. plot(tempData)

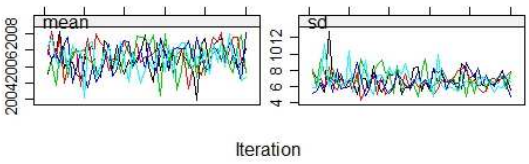

Figure 4: The Result of Mean and SD

IMPLEMENTATION OF K MEANS ALGORITHM WITH THE HELP OF DATASET

The enormous data collected from production company is analyzed using R. It explains how big data is analyzed with R studio and linking big data with R tool, so the time consumed will be reduced. The large volume of data will be clustered and analyzed to reduce the contents of data in k mean cluster the data will be reduced [14].

Algorithm K-Means (Thick data)

{S -input data set csv

M-matrix format

C-Cluster using K-Means}

1. Take the input S
2. M<-data. atrix(S)

   //convert to matrix format
3. C<-k means(M,8,10)

   //clusters using k-means cluster (8X10).
4. cluster<-C$ cluster//save the cluster into cluster
5. plot(cluster) // plotting the graph of the clustered data.
6. Summary (cluster)

Table 3: Summary of Cluster

Min.	1st Qu.	Median	Mean	3rd Qu	Max
1.000	3.000	5.000	4.596	6.000	8.000

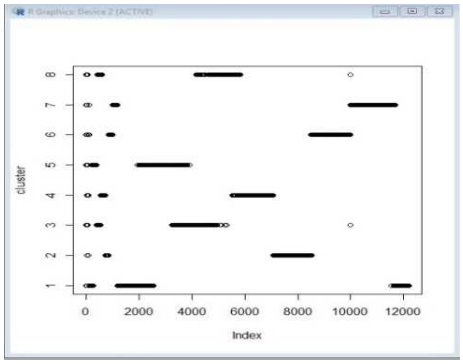

Figure 5: The Results of the Big Data After k-means Clustering

RESULT OF K-MEANS CLUSTER

This K-means algorithm is used for the huge date which is complete. The classification algorithm will differentiate the data which is complete and the missing data. Those data which is complete will be sent to K-means cluster to analyze the data and the result will be plotted with the help of graph. Figure 5 is the result of the data which is complete and clustered.

IMPLEMENTATION OF EM ALGORITHM WITH THE HELP OF TRAINING DATASET

The Expectation-Maximization (EM) algorithm is used to cluster huge datasets and high dimensional. It improves the quality of the solution. It is simple and can be handled in real-world application.

Algorithm Expectation Maximization

{x1 -input data set csv

x2-list which contains μ, π

ret-Cluster using EM

assign. lass-assigning Class ID

nclass-assign number of classes}

1. Load EM cluster package
2. Take the input x1
3. x2 <- simple. nit(x1, nclass = 10)

   //the list mainly contains pi, u and LTsigma which returned from initial.
4. x2 <- shortemcluster(x1, x2)

   summary(x2)

   //clustering the objects
5. R <- emcluster(x1, x2, assign. lass = TRUE)

   summary(R) //assigning it to class id
6. R.1 <- starts. ia. vd(x1, nclass = 10, method = "em")

   //desired number of clusters are assigned
7. summary(R.1)
8. plotem(R.1, x1)
9. plotem(R, x1)
10. plotem(x2, x1)

    // plotting em graph of the clustered data.

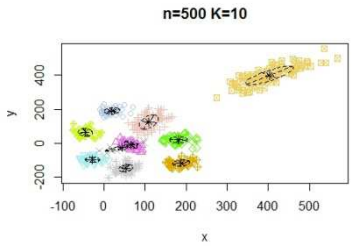

Figure 6: The results of the Big Data after EM Algorithm for Clustering

RESULT OF EM ALGORITHM

The EM algorithm is used for analyzing the missing data, where it will take the data and if will randomly fills the values to the missing data and the data will be analyzed. Figure 6 is the result of the data which is analyzed for the data making them into clusters and for each cluster, a number is assigned and the graph is plotted for those values.

Summary (Cluster)

Table 4: Summary of EM Cluster

n = 500, p = 2, nclass = 10, flag = 0, total parameters = 59.					
nc:	40	29	37	44	25
pi:	0.084	0.198	0.070	0.076	0.098
mean:	-27.70 -96.17	403.1 398.9	45.57 62.66	187.8 -118.6	63.3 -11.3
nc:	56	169	29	55	16
pi:	0.080	0.114	0.102	0.088	0.090
mean:	19.39 189.64	180.91 19.08	52.82 -148.39	96.65 91.05	122.0 159.8

Performance Improvement

With the efficient integration of the analytical model, it is observed that the performance of the Hadoop parallel cluster has improved, performance is monitored with the help of ganglia monitoring tool and results are shown in terms of improvement with respect to CPU utilization in the system.

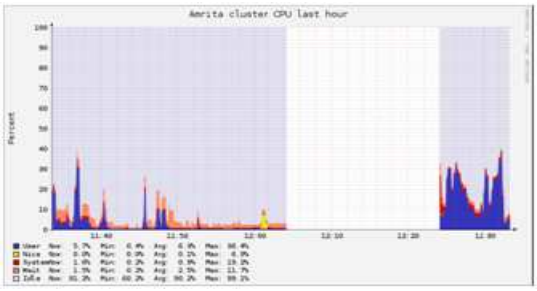

Figure 7: CPU Utilization

figure 7 depicts variations and improvements in the CPU utilization of proposed Cluster. Thus, an overall reduction of load, improvement of CPU utilization has helped in optimizing the performance. Thus, the proposed system is an ideal solution for real- time applications like monitoring of the production data in large companies.

CONCLUSIONS

The research carried out enhances the analytic part of Big data handling in Hadoop platform for production companies. Initial categorization helps in bifurcating data to sparse and thick and then based on the data type using a multi -model Hadoop analytic with K-Mean and EM is performed which increases the overall performance of Big data analytics and processing, The system can be improvised by introducing more categorization algorithms and analytic algorithms which can work parallelly to improve a production companies.

REFERENCES

1. *A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," Proc. VLDB Endow., vol. 5, no. 12, pp. 2032–2033, 2012.*
2. *K. Abarna, M. Rajamani, and S. K. Vasudevan, "Big data analytics: A detailed gaze and a technical review," Int. J. Appl. Eng. Res., vol. 9, no. 11, pp. 1735–1751, 2014.*
3. *C. Ji, Y. Li, W. Qiu, Y. Jin, Y. Xu, U. Awada, K. Li, and W. Qu, "Big Data Processing: Big Challenges And Opportunities," J. Interconnect. Networks, vol. 13, no. 3/4, pp. 1–19, 2012.*
4. *A. Mukherjee, J. Datta, R. Jorapur, R. Singhvi, S. Haloi, and W. Akram, "Shared disk big data analytics with Apache Hadoop," in 2012 19th International Conference on High Performance Computing, HiPC 2012, 2012.*
5. *M. G. M. Mohan, S. K. Augustin, and V. S. K. Roshni, "A BigData approach for classification and prediction of student result using MapReduce," in 2015 IEEE Recent Advances in Intelligent Computational Systems, RAICS 2015, 2016, pp. 145–150.*
6. *A. B. Patel, M. Birla, and U. Nair, "Addressing big data problem using Hadoop and Map Reduce," in 3rd Nirma University International Conference on Engineering, NUiCONE 2012, 2012.*
7. *B. Li, E. Mazur, Y. Diao, A. McGregor, and P. Shenoy, "A platform for scalable one-pass analytics using MapReduce," Proc. 2011 Int. Conf. Manag. data - SIGMOD '11, p. 985, 2011.*
8. *J. Han, Y. Liu, and X. Sun, "A scalable random forest algorithm based on MapReduce," in Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS, 2013, pp. 849–852.*
9. *B. Mandal, S. Sethi, and R. K. Sahoo, "Architecture of efficient word processing using Hadoop MapReduce for big data applications," in Proceedings - 2015 International Conference on Man and Machine Interfacing, MAMI 2015, 2016.*
10. *I. Taleb, R. Dssouli, and M. A. Serhani, "Big Data Pre-processing: A Quality Framework," 2015 IEEE Int. Congr. Big Data, pp. 191–198, 2015.*
11. *J. Dittrich and J. Quian, "Efficient Big Data Processing in Hadoop MapReduce," Proc. VLDB Endow., vol. 5, no. 12, pp. 2014–2015, 2012.*
12. *Chandrashekar, S. V. "Inference of feminist contemplation's in reduction of women victimization." BEST: International Journal of Humanities, Arts, Medicine and Sciences 4.12 (2016): 63-70.*
13. *C. Ji, Y. Li, W. Qiu, U. Awada, and K. Li, "Big data processing in cloud computing environments," Proc. 2012 Int. Symp. Pervasive Syst. Algorithms, Networks, I-SPAN 2012, pp. 17–23, 2012.*
14. *Jyothi, B. Sai, and S. Jyothi. "A study on Big Data Modeling Techniques." International Journal of Computer Networking, Wireless And Mobile Communications (IJCNWMC) 5 (2015): 2250-1568.*
15. *O. D. L. Yejas, W. Zhuang, and A. Pannu, "Big R: Large-scale analytics on hadoop using R," in Proceedings - 2014 IEEE International Congress on Big Data, BigData Congress 2014, 2014, pp. 570–577.*
16. *"A Study on Evolution of Data Anal Ytics To Big Data Anal Ytics and Its Research Scope," 2015.*
17. *R. Huang and W. Xu, "Performance evaluation of enabling logistic regression for big data with R," Big Data (Big Data), 2015 IEEE Int. Conf., no. October 2015, pp. 2517–2524, 2015.*